HIERARCHICAL AGGLOMERATIVE CLUSTER ANALYSIS WITH A CONTIGUITY CONSTRAINT

by

Brant H. Wipperman
B. Sc., Simon Fraser University, 1999

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
January 2004

Approval

Name: Brant H. Wipperman

Degree: Master of Science

Title of Project: Hierarchical Agglomerative Cluster Analysis with a Contiguity Constraint

Examining Committee:

Chair: **Dr. Boxin Tang**
Associate Professor

Dr. Michael A. Stephens
Senior Supervisor

Dr. Richard A. Lockhart
Simon Fraser University
Professor

Dr. Gary Parker
External Examiner
Simon Fraser University

Date Approved: January 22, 2004

PARTIAL COPYRIGHT LICENCE

Title of Thesis/Project/Extended Essay:

Hierarchical Agglomerative Cluster Analysis with a Contiguity Constraint

Author:

(Signature)

Brant Wipperman
(Name)

January 19, 2004
(Date Signed)

Abstract

Cluster analysis is a technique for finding group structure in data; it is a branch of multivariate statistics which has been applied in many disciplines. The most common method of cluster analysis is hierarchical agglomeration. Several algorithms are discussed, with a focus on complete linkage. Constrained classification is then presented, specifically the case in which members of a cluster are required to be geographically contiguous. An example is provided, illustrating the creation of territories for automobile insurance in British Columbia, Canada. The dissimilarities between objects are measured by symmetrized deviance drops. This approach may be described as model-based clustering subject to contiguity constraints.

Acknowledgements

One does not survive a full decade of university without a myriad of support. My list may be longer than some others due to the extra efforts which have made this degree possible.

First of all, I would like to thank my supervisors. It seems fitting to acknowledge Michael Stephens in a list format. I am grateful for his following contributions: (a) friendship; (b) house-sitting opportunities; (c) generous research assistantships; (d) embracing my project idea; (e) enduring numerous Microsoft Word glitches; (f) tolerating my German writing style, with its frequently recurring adjectives; (g) patience with my work schedule. Richard Lockhart has been personally responsible for more than half of my Masters credits. He was extremely understanding of my situation and offered all the help that I needed to finish the courses successfully.

I would also like to mention a few other members of the Simon Fraser University Statistics community: Egon Simons for his support in the co-op program; Norman Reilly and Ken Collins for keeping the actuarial program alive; Randy Sitter for recruiting me to the Masters program; Larry Weldon for some interesting TA experiences; Sylvia Holmes for processing endless on-leave and part-time forms; Gary Parker for acting as my external examiner; and too many graduate students to mention for their assistance over the years. It is a credit to the kindness of the department that I have been able to complete this degree while working full-time.

A number of past and present employees of the Insurance Corporation of British Columbia were instrumental in this project. I would like to thank Christian Fournier, former Chief Actuary, for all his encouragement and oversight of this project. Also to Michael Chan and Sandra Creaney, who hired me and supervised parts of the project. Thank you to Harry Pylman, Chief Underwriter, and Bill Weiland, Actuarial Consultant, for taking the time to read my drafts. I was extremely thankful for the use of a laptop and that I was allowed time off whenever I needed it.

I wish to thank my wife, Sara Wipperman, for all her understanding and encouragement during this long process. My parents, Paul and Sheila Wipperman, have been constant cheerleaders throughout my various academic endeavours. It is a blessing to have a second set of supportive parents, Dietrich and Jeanette Wittkowski.

The members of the Grace Community Church in Pitt Meadows have been very special to me. Thank you for your continued prayers, friendship, caring and meals.

I give all the credit for this project to my Lord, Jesus Christ. May I live a life worthy of the calling I have received.

Table of Contents

List of Figures

List of Tables

Glossary

This glossary defines vocabulary common to automobile insurance rating in British Columbia.

claims rated scale (CRS)	an individual's driving record and experience as described by an integral value; new drivers begin at level 0 and receive minus 1 credit for each year of accident-free driving
collision coverage	a product protecting motorists against property damage when their vehicle is determined to be at-fault in a crash
comprehensive (comp.) coverage	a product protecting motorists against property damage to vehicles resulting from theft, vandalism, fire, animal collision, glass breakage and other perils
credibility	a measure of the amount of trust placed in the precision of a given estimate; related to confidence intervals
deductible	the dollar amount of a loss which is retained, that is, paid by the insured
exposure	a measure of baseline risk to an insurer; one exposure unit is a policy with a 12 month term, a 6 month policy would count as half an exposure unit
frequency (freq.)	the number of claims made per exposure; usually expressed as a percentage
Insurance Corporation of British Columbia (ICBC)	a crown corporation established in 1973 to provide automobile insurance coverage to British Columbia motorists

lessee	a person leasing a vehicle from its owner
lessor	the owner of a vehicle which is leased
loss cost	the average dollar amount of claims per exposure; equivalently, the product of frequency and severity
loss ratio	the percentage of premium collected which is used to pay for claims
mandatory insurance	the minimum amount of coverage required; provided by ICBC
policy	a contract specifying how an insurer will compensate an insured for losses arising from certain events
private passenger	a vehicle used primarily for pleasure, commuting, or business; excludes motor homes, motorcycles, collector vehicles, trailers, buses, taxis, limousines, etc.
rate class	a description of vehicle use and type represented by a three digit code; may incorporate such information as distance driven, vehicle weight, engine size, and passenger capacity
rate group	a vehicle rating assigned to specific vehicle make, model and model year combinations; based on factors such as repair cost and theft frequency
RoadStar	a designation used for ICBC's best customers; CRS level -9 or better
severity	the dollar amount of an insurance claim; may also refer to an average amount
short-term	a contract written for less than a full 12 months
term	the period for which an insurance policy is in effect; usually described by the difference between the effective and expiry dates of the policy
territory	a geographical region used for insurance rating

1 Introduction

1.1 Project History

The Insurance Corporation of British Columbia (ICBC) uses a set of 14 geographic territories to help set provincial automobile insurance rates. A map of these territories is provided in Figure 1.1.

© 1991 Insurance Corporation of British Columbia, by permission

Figure 1.1 Map of ICBC territories

Urban sprawl and a multi-year rate freeze have contributed to a large increase in the variation of loss ratios within these territories. This is especially true in Greater Vancouver and the Fraser Valley, which are currently divided into only three territories, but account for more than half of the company's policies (see Figure 1.2). One of these territories contains nearly half of the province's vehicles alone.

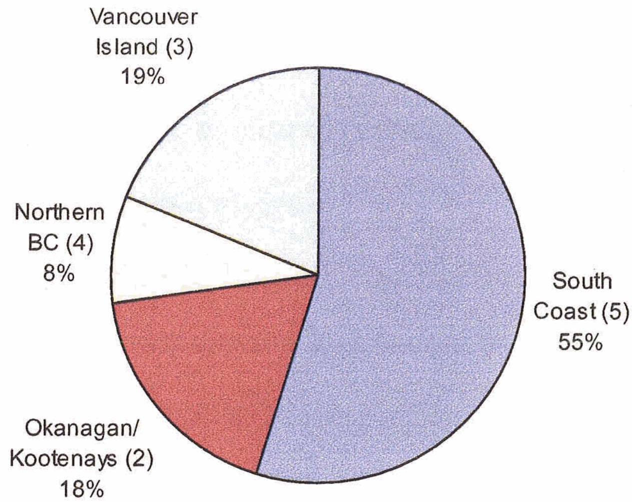

Figure 1.2 Regional distribution of policies
(number of territories in parentheses)

The inhomogeneity within territories has created an excellent opportunity for ICBC's competitors. They are able to offer substantially lower rates to customers in those regions where rates do not accurately reflect the level of risk. Even in the absence of competition on mandatory insurance, there was a social concern arising from this large imbalance.

The aim of the cluster analysis which follows was to identify geographically contiguous regions with similar claims experience, which could form a greater number of smaller and more efficient territories. This was accomplished by studying various attributes of the claims histories of existing territories.

1.2 Data Objects

The original region of interest for the study comprised an area extending from West Vancouver to Chilliwack. This was later extended to include the rest of British Columbia. In a cluster analysis, the objects to be clustered must be clearly defined. In this project, the objects are geographical areas (polygons).

The following municipalities were deemed large enough to identify one or more objects: Abbotsford/Chilliwack, Campbell River, Castlegar/Nelson/Trail, Courtenay/Comox, Cranbrook/Kimberley, Dawson Creek/Fort St. John, Duncan, Kamloops, Greater Kelowna, Merritt, Mission, Nanaimo/Ladysmith, Parksville/Qualicum Beach, Penticton, Port Alberni, Powell River, Prince George/Quesnel/Williams Lake, Prince Rupert/ Terrace/Kitimat, Salmon Arm, Greater Vancouver, Vernon, and Greater Victoria.

Outside of these urban centres, it was more challenging to define the objects, as large enough objects were not available as portions of municipalities. These other regions are typically a mixture of small towns and rural areas. These were combined into credible objects based on geographical proximity and claim counts.

There was a total of 247 objects in the province's 14 territories. The number of objects in each territory generally ranged from two to 36, with only the Lower Mainland having more. Territories with six or fewer objects were not clustered (see section 3.10).

Some objects had very little claims experience over the study period, due to low exposure counts. These smaller objects often did not merge until late in the clustering process. The rural objects tended to be smaller in terms of claims and policy counts than the urban objects. In general, rural areas tended to have lower crash rates than the Lower Mainland but other types of claims, such as for glass damage, were more frequent. Standards for full credibility of objects were defined by a minimum number of claims for collision and comprehensive. These minimum claims standards were relaxed somewhat for rural objects, and other data such as population and policy counts in force were considered.

Instead of discarding the data, the claims from smaller objects were pre-merged with neighbouring regions based on a comparison of the mean frequencies and severities. A total of 15 objects were pre-merged, nine in urban centres and six in rural areas. This left a total of 232 objects for the entire province.

Carvalho et al. (1996) describe a more formal procedure for aggregation, as they call it. This creates objects which exceed a pre-specified minimum population.

1.3 Data Attributes

Two insurance coverages, collision and comprehensive, were examined for private passenger motor vehicles over a five year period from 1997-2001. The study of property and casualty insurance claims is generally split into two components, frequency and severity, which are modelled separately (Klugman et al., 1998). For each object, the severity of all claims was captured. No outliers were observed, with all amounts falling below $100,000.

It is possible to determine how many claims have been made during the term of a policy, given that at least one claim has occurred. Claims records with the same license plate and expiry date belong to the same policy term. The highest number of claims observed in one policy term was five for collision and nine for comprehensive.

Exposure was calculated as the difference between the expiry date and the inception or renewal date. However, there were some complications. Policy cancellations are not reflected in an earlier expiry date, so exposure may be overstated for vehicles which are written off. That is, we essentially assume that no further accidents would have occurred had the vehicle still been on the road. The experience from policies which overlapped either end of the five year period was censored to be consistent with the scope of the study.

It is not possible to gather exposure data directly on policies without claims. Instead, we obtained total object exposures for the period. Then the total exposure of claims-free policies was given by the difference between that of all policies, and those with at least one claim.

1.4 Data Quality

All aspects of geographical information for policies and claims were required to be consistent. This resulted in the deletion of about 1.5% of claims and 1.7% of policy records. There was also a small number of coding errors in the claims database, comprising 0.1% of records. These claims were omitted since they could not be accurately assigned to an object.

There was a problem with claims on leased vehicles because the geographical information available is that of the lessor, not the lessee. Leased vehicles accounted for about 10% of claims and have been removed from the analysis because they would have biased the results near major leasing locations. It would have been unfair to charge the actual residents based on the experience of these leased vehicles. In general, leased vehicles tended to have higher frequency and severity characteristics. Loss ratios were especially higher for collision coverage.

2 Survey of Cluster Analysis

2.1 Introduction to Cluster Analysis

The general problem of cluster analysis can be formulated as the partitioning of a set of N objects into K clusters. Partitioning is based on data collected from each of the objects. The data may be univariate, or describe multiple features, potentially with different scales of measurement. Dissimilarity is a mathematical measure of distance between a pair of objects. It may use the original data values, some estimated parameters, or the likelihoods of statistical models.

Initially, all clusters consist of single objects. Later, clusters may contain one or more objects. The dissimilarity between clusters is calculated from the dissimilarities of pairs of objects, one object from each cluster. At each step in the clustering process, the two clusters with the lowest dissimilarity are joined, subject to any constraints.

The many applications of cluster analysis are too numerous to list here and vary greatly. Taxonomy, marketing, epidemiology, chemistry and library science are a few of the subjects mentioned in the literature. The respective objects to cluster in these disciplines might be species, customer groups, health regions, elements and books.

Hartigan (1975), Mirkin (1996) and Gordon (1999) each describe several purposes served by a cluster analysis:

(i) naming of objects in a way that distinguishes them from others;

(ii) summarization and simplification of data so that characteristics of clusters, rather than individual objects, can be studied;

(iii) convenient display of information;

(iv) use of resulting groupings for prediction;

(v) inspiration to create hypotheses and theories.

In the creation of territories for insurance rating, the objective is best described by purpose (iv) above. The clusters of objects formed will become part of the rating structure, and a major portion of an insurance rate will be a prediction of the expected value of claims in clusters.

Data for cluster analysis may be on an interval, ratio, ordinal, or categorical scale (Hartigan, 1975). There are three main ways to handle multiple data types (Romesburg, 1984):

(i) the scale differences can be ignored;

(ii) multiple analyses could be performed on the different variables;

(iii) the continuous attributes may be summarized into counts in intervals.

The second approach will be employed in this project to deal with counts and continuous data.

Standardization is also an important consideration as there is a potential that a subset of the variables could dominate the dissimilarity measure. It is important to weigh the variables carefully so if standardization is not used, simpler techniques such as omission of variables or repetition of data should be employed (Romesburg, 1984).

There are two major objectives to consider which affect the mechanics of how clusters are formed. These goals are homogeneity within the resulting groups and differentiation between them. The balance between these considerations is a determining factor in choosing the clustering method.

2.2 *Methods of Cluster Analysis*

Clustering methods are of two distinct types: hierarchical and non-hierarchical. Non-hierarchical methods take an existing classification and re-assign the membership of objects. Hierarchical methods reveal the cluster membership of objects for each possible number of clusters, providing a complete picture of group structure. They are by far the most popular methods in practice and will be utilized in chapter 3. Hierarchical methods can be further split by the direction of clustering. Agglomerative methods start from the bottom-up with N singleton objects and join pairs of objects or pairs of clusters until all objects inhabit a single cluster. Conversely, divisive methods operate from the top-down by choosing a cluster at each step, and splitting it.

Hierarchical agglomerative algorithms may be grouped into three main categories: linkage, variance, and centroid. Linkage methods include single link (also known as minimum link or nearest neighbour), complete link (also known as maximum link or further neighbour), average link, and weighted average link. The other types of algorithms, variance and centroid, use Euclidean distance as the measure of dissimilarity (Gordon, 1999).

Hierarchical agglomeration and other methods of cluster analysis generally require the specification of a square dissimilarity matrix $\mathbf{D}$ with entries D_{ij}, which represent the dissimilarities between the data from each pair of objects, O_i and O_j. The elements of $\mathbf{D}$ are subject to the following three conditions (Gordon, 1996a):

(i) $D_{ij} \geq 0$;

(ii) $D_{ii} = 0$;

(iii) $D_{ij} = D_{ji}$.

In single linkage, the clusters with the two most similar member objects are always joined. In complete linkage, the dissimilarity between two clusters is measured by the maximum dissimilarity between member objects. Clusters are then formed by joining those with the lowest dissimilarities. Single linkage often results in well-differentiated groups but is subject to chaining. This phenomenon occurs when one of the clusters continually grows. This is because there are many comparisons with this large cluster, and one of the dissimilarities is likely to result in another object joining it. Complete linkage can be thought of as a trial-and-error approach as every possible merger is analyzed in order to minimize the dissimilarity within clusters. The susceptibility to chaining disappears and clusters tend to be cohesive and of similar size. A drawback of complete linkage is that distinct clusters might be quite similar, but this is not a concern here. Finally, average and weighted average linkage seek to balance the two objectives of homogeneity and differentiation (Gordon, 1987) by looking at average differences between groups. Complete linkage is used for the analyses to follow in chapter 3.

2.3 Constrained Classification

One complication in cluster analysis arises when constraints exist on the cluster membership of objects. DeSarbo and Mahajan (1984) attribute the introduction of the topic of constrained cluster analysis to Gordon (1973). Constrained analyses may sometimes be performed in order to allow comparisons with unconstrained analyses, and are often easier to interpret (Gordon, 1996b). Geographical contiguity is the constraint which most often affects spatial cluster analyses. Another common constraint is on the number of objects in a cluster. This type of constraint can also be addressed outside of the analysis by simply choosing a partition with more clusters (Murtagh, 1985). Another frequently occurring constraint in cluster analysis is on the specific composition of classes.

The sources of constraints may be dichotomized as inherent or imposed conditions (Murtagh, 1985). Inherent, or internal reasons have to do with the physical resemblances of objects (DeSarbo and Mahajan, 1984). Imposed, or external constraints may be due to policy or resources (Gordon, 1996b). In this project, both types of constraints are present. The fact that territories are required to be contiguous may be regarded as an inherent condition, while the restriction that the new territories preserve the boundaries of the existing territories is an imposed legal constraint.

Contiguity constraints tend to lead to the problem of single objects never joining a cluster, because an object may have a small number of neighbours, and may happen to be dissimilar to all of them. On the other hand, constrained analyses are not as dependent on the chosen clustering method, since many fewer solutions are possible (Gordon, 1996a).

Contiguity constraints are usually handled in one of three ways. First, they can be ignored altogether in the clustering and only assessed afterwards upon inspection of a map (Gordon, 1996b). However, this can be misleading as contiguous objects may not be similar and spatially varying effects could be overlooked. A second strategy sometimes employed is to quantify contiguity as part of the dissimilarity measure (Murtagh, 1985). This involves adding a term to each entry of the dissimilarity matrix which describes the geographical distance between the objects. But this is very subjective as results will depend on the magnitude of the penalty assessed to non-contiguous objects.

The third approach is to utilize a contiguity matrix (Gordon, 1996b), which must be consulted before any merger is allowed. There are sometimes difficulties in defining contiguity exactly, and the matrix must be specified such that there is a path between any two objects. A disjoint set is a group of objects with a path connecting them but where none of the objects are contiguous to any object outside the set. The disjoint sets of objects could be analyzed separately. If objects are arranged in a square grid, such as plots of land, an object may have four neighbours if common edges are required for contiguity, or eight neighbours if intersection at a corner is sufficient for contiguity (Murtagh, 1985). If the objects are points, contiguity may be defined as being within a certain radius, or being one of the nearest neighbours (Murtagh, 1985). Gordon (1999) also points out that neighbouring objects are not necessarily contiguous, e.g. if they are separated by mountain ranges, bodies of water, etc.

The contiguity matrix $\mathbf{C}$ is ordinarily a square, symmetric matrix with binary entries, C_{ij}. Murtagh (1985) suggests that a continuous measure could also be used, but in this case binary values would seem to suffice. Let $C_{ij} = 1$ if objects O_i and O_j are contiguous, and $C_{ij} = 0$ otherwise. The diagonal entries of the matrix, C_{ii}, are not meaningful and if we also assume that $C_{ij} = C_{ji}$, then only the sub-diagonal elements of $\mathbf{C}$ need to be specified. Overall, the challenges presented by using a contiguity matrix seemed the least difficult to overcome, so this was selected as the way to deal with the constraint of geographical contiguity. More details on the specification of the contiguity matrix will be provided in section 3.2.

2.4 Cluster Validation

A cluster analysis always produces a partition, even if there are no truly different groups present in the data (Stockburger, 1996). However, the presence of heterogeneity is rarely tested in practice (Gordon, 1999). Tests of a given hierarchy are focused on three main areas: homogeneity, differentiation and stability. Various functions of dissimilarities within and outside of clusters may be used to assess homogeneity (Gordon, 1999). Differences between the times of original cluster formation and eventual amalgamation can measure the differentiation between groups (Gordon, 1996a). One way to determine the stability of a hierarchy is to run the analysis on several data sets or with separate variables, or to use multiple clustering methods, and then compare the various results (Gordon, 1999). A common stopping rule used to choose the number of clusters is to look at a tree diagram and determine at which step, say step k, there is a large gap in the dissimilarity values of merged clusters (Wulder, 2002). This implies that the previous merger, at step k-1, combined much more similar clusters than those merged at step k.

The theory behind post-analysis evaluation is not a well-developed field and informal methods are common, such as an assessment of whether the project goal was achieved within an acceptable level of tolerance (Romesburg, 1984). This assessment may be based mainly on intuition or prior beliefs.

3 Application - Insurance Claims Data

3.1 Outline of Analysis

This chapter describes the process of clustering the objects, namely the regions as described in section 1.2, in a single territory. Suppose the territory contains n objects, O_k, k = 1...n. There are four sets of data to be used in clustering: collision severity, comprehensive severity, collision frequency and comprehensive frequency, as discussed in section 1.3. The main complication in the analyses is a constraint based on geographical contiguity, a topic which was introduced in section 2.3. A numerical example of clustering with collision severity data will be shown in sections 3.6 and following.

Romesburg (1984) and Gordon (1999) each outline six major steps in a cluster analysis, which are largely applicable here. The first three steps involve selection of the objects, O_k, and variable(s) of interest, followed by standardization of the data, if necessary. The next two steps are to define a measure of dissimilarity, $\mathbf{D}$, among objects and choose a clustering method. Finally, in the presentation step, the number of clusters is determined, results are interpreted and significance may be tested. In model-based clustering, it is also necessary to choose a distributional model for the variable(s). For this project, a description of these steps is provided in Table 3.1.

Step	Description	Sections
Objects	Geographical areas, e.g. municipalities	1.2
Variables	Collision severity, Comp. severity, Collision freq., Comp. freq.	1.3
Standardization	Rate Group and Deductible, or Rate Class and CRS factors	3.3, 3.4
Model	Lognormal, Poisson, or Negative Binomial distributions	3.5, 3.7
Dissimilarity	Symmetrized deviance drops	3.5
Method	Complete linkage	2.2, 3.5
Presentation	Consensus rules and outlier re-allocation	3.8, 3.9

Table 3.1 Steps in a cluster analysis

The approach employed here may be summarized as hierarchical agglomerative classification, using the complete linkage clustering method, and subject to contiguity constraints. The output is a hierarchical arrangement of clusters joined at increasing levels of dissimilarity. The clustering is run separately on each of the four combinations of coverage (collision, comprehensive) and attribute (frequency, severity). But since the desired output is a single set of new territories, the results need to be combined in some way.

3.2 Contiguity Matrix

Suppose the contiguity matrix $\mathbf{C}$ is a symmetric n x n matrix with elements C_{ij}, indexed by the data objects. The only objects allowed to merge are those which are geographically adjacent, that is, pairs of objects O_i and O_j with $C_{ij} = 1$. $\mathbf{C}$ is updated and utilized in each step of the clustering algorithm. For instance, after step k, the updated contiguity matrix will be denoted by $\mathbf{C}^{n-k}$.

Contiguity must be precisely defined as there are many possible ambiguities. The most common criterion for two objects to be contiguous is a common land boundary of reasonable length. This means that the boundaries do not simply meet at a single corner, or for a small number of city blocks. In rural areas, it is also a requirement that at least one road crosses the common boundary. Objects will also be considered neighbours if they are separated by a body of water but accessible by a bridge, tunnel or vehicle ferry. In fact, it is necessary to define objects connected in this way as contiguous because otherwise certain groups of objects might never be able to merge with any others. The city of Richmond is an example since access to neighbouring municipalities requires travelling across a bridge or through a tunnel.

Applying these rules consistently across the province results in an average of two to four neighbours per object, depending on the territory under consideration. For individual objects, the number of neighbours was as low as one for objects on a territory boundary. The largest number of neighbours observed for a single object was seven.

3.3 Data Standardization

In order to isolate the effect of territory from all other rating criteria, the frequency and severity data had to be adjusted for differences due to rate class, claims rated scale, rate group, and deductible. It was decided to adjust each combination of coverage and attribute for two variables, as shown in Table 3.2. These were determined to be rate group and deductible for all combinations other than collision frequency, which were standardized for rate class and claims rated scale differences.

Coverage/Attribute	Variables chosen	Explanation for chosen variables
Collision severity	Rate group, deductible	Repair costs, speed of vehicles
Comp. severity	Rate group, deductible	Repair costs, many small claims
Collision frequency	Rate class, CRS	Distance driven, driving history/experience
Comp. frequency	Rate group, deductible	New vehicles targeted, small claims eliminated

Table 3.2 Standardization of claims data

To ensure enough occurrences of each bivariate combination, rating variables were grouped into intervals as shown in Tables 3.3 - 3.6.

Interval	Rate Class	Vehicle Use
1	001	Pleasure use only
2	002	To and from work or school
3	003, 004	Commuting under 15 km, or park and ride
4	005	Seniors (pleasure only)
5	007	Business use
6	021 - 027	Experienced drivers only (licensed 10+ years)

Table 3.3 Rate class intervals

Interval	CRS Level	Status
1	$\leq$ -15	RoadStar Gold
2	-9 to -14	RoadStar
3	$\geq$ -8	Non-RoadStar

Table 3.4 Claims rated scale (CRS) intervals

Interval	Rate Group
1	00-02
2	03-05
3	06-08
4	09-11
5	12-14
6	15-23

Table 3.5 Rate group intervals

Interval	Deductible
1	≤ $300
2	≥ $500

Table 3.6 Deductible intervals

There are 11 rate classes which were grouped into six categories designed to preserve similarity of vehicle uses, while aggregating newer or less frequently used classes. The three claim rated scale intervals are in common use and distribute policies into roughly equal proportions. The main goal of collapsing the 24 rate groups into six intervals was to obtain similarly sized groups of policies. Minimum deductibles have historically ranged from $100 to $300, so the $500 and up interval was chosen to separate those customers who intentionally selected a higher deductible.

There is a substantial amount of variation among data objects in the distribution of the above variables; otherwise, of course, standardization would not be necessary. For example, wealthier areas tended to have more vehicles with high rate groups and a greater proportion of policies with higher deductibles. There were also large differences noted in the proportions of non-RoadStars and experienced drivers.

3.4 Adjustment Factors

There are 12 possible combinations of the 6 rate group intervals and 2 deductible intervals. There are 18 possible combinations of the 6 rate class intervals and 3 claims rated scale intervals. Since the example to follow uses collision severity for the clustering, we suppose that the data are being adjusted for rate group intervals, indexed by s = 1...6, and deductible intervals, indexed by t = 1...2. For each object O_k within a territory, the proportion p_{stk} of the vehicles belonging to each rate group/deductible combination was calculated. The proportions p_{st} over all objects were also determined by allocating all vehicles in the territory to the appropriate combinations.

Similarly, average frequencies f_{st} and severities z_{st} were calculated within the entire territory for all rate group/deductible combinations. These were then multiplied by the object proportions p_{stk} and summed to obtain an expected frequency f_k and severity z_k for each object O_k, k = 1...n:

$$f_k = \sum_{s=1}^{6} \sum_{t=1}^{2} p_{stk}\, f_{st}$$

$$z_k = \sum_{s=1}^{6} \sum_{t=1}^{2} p_{stk}\, z_{st}$$

These values were compared to an expected frequency $\bar{f}$ and severity $\bar{z}$ for the entire territory, which are based on the overall proportions p_{st}:

$$\bar{f} = \sum_{s=1}^{6} \sum_{t=1}^{2} p_{st} f_{st}$$

$$\bar{z} = \sum_{s=1}^{6} \sum_{t=1}^{2} p_{st} z_{st}$$

Then for each object O_k, $k = 1 \ldots n$, an adjustment factor was calculated for both frequency (factor u_k) and severity (factor v_k):

$$u_k = \bar{f} / f_k$$

$$v_k = \bar{z} / z_k$$

The values of these adjustment factors fell mostly between 0.95 and 1.05, with almost all the values between 0.90 and 1.10. It was not considered worthwhile to use a more complicated method of standardization.

3.5 Methodology for Severity Data

This section describes the models, likelihoods and dissimilarities for severity data, both collision and comprehensive. Frequency data are discussed in section 3.7. The severity data for an object O_k consists of a list of individual claim amounts y_{kl}, $l = 1 \ldots m_k$. The application of the adjustment factors was straightforward for severity. Each individual claim amount y_{kl} was multiplied by v_k, but the adjusted severities will still be referred to as y_{kl}, $l = 1 \ldots m_k$. A lognormal model was fit to these adjusted severities. The density function of the lognormal distribution has the form:

$$f\,(y; \mu, \sigma^2) = \frac{\exp\{-[\log(y) - \mu]^2 / 2\sigma^2\}}{\sqrt{2\pi}\,\sigma\, y} \qquad\qquad y > 0$$

For object O_k, the maximum likelihood estimates for the parameters μ_k and σ_k^2 of the lognormal model are as follows:

$$\hat{\mu}_k = \bar{x}_k = [\textstyle\sum_{l=1}^{m_k} \log(y_{kl})] / m_k \qquad\qquad k \in \{1, \ldots, n\}$$

$$\hat{\sigma}_k{}^2 = s_k{}^2 = \{\textstyle\sum_{l=1}^{m_k} [\log(y_{kl})]^2\} / m_k - \bar{x}_k{}^2 \qquad\qquad k \in \{1, \ldots, n\}$$

3.5.1 Dissimilarity Measures

We can now construct the log-likelihood matrix $\mathbf{L}$. The entries L_{ij} are given by the log-likelihoods of the object O_j data under the models estimated with the object O_i data:

$$L_{ij} = \log \textstyle\prod_{l=1}^{m_j} f\,(y_{jl}; \hat{\mu}_i, \hat{\sigma}_i{}^2) = -\sum_{l=1}^{m_j} \{[\log(y_{jl}) - \hat{\mu}_i]^2 / 2\,\hat{\sigma}_i{}^2\} - m_j \log(\sqrt{2\pi}\,\hat{\sigma}_i) - \sum_{l=1}^{m_j} \log(y_{jl})$$

We could now measure the dissimilarity between a pair of objects O_i and O_j with the symmetrized distance (Smyth, 1997):

$$D^{L}_{ij} = (L_{ij} + L_{ji}) / 2 \qquad i, j = 1 \ldots n; \; j < i$$

Note the building blocks of the dissimilarities are model likelihoods, not simply data or parameters. Model-based clustering uses the comparison of statistical models to assess dissimilarity. However, Smyth's original clusters were all the same size, and this is not the case here. Likelihoods depend on the number of data points, and the rows of the log-likelihood matrix $\mathbf{L}$ tend to be very similar. That is, the spread among log-likelihoods, L_{ik}, $i = 1 \ldots n$, involving the data from object O_k, is small, as only the parameter estimates, $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, are varying. If we proceeded as above, the objects with fewer claims might merge sooner than they should, because there is not enough information to conclude that they are dissimilar.

One might think of averaging log-likelihoods per observation but empirically this turns out to be unsatisfactory. For example, using the Lower Mainland data we would have expected a priori that the most similar pair of objects would be the two listed in Table 3.7 below, since their means and standard deviations modelled by the lognormal distribution differed by only a dollar or two.

Object (O_k)	Claim Count (m_k)	Mean	Standard Deviation	Mode
O_{36}	1,475	4,086.64	9,284.45	427.46
O_{12}	3,316	4,087.48	9,286.88	427.51

Table 3.7 Estimated parameters of lognormal collision severity models

But when the dissimilarities were constructed from the average of the two log-likelihoods per data point, these two objects were not the most similar. Another possibility would be to use log-likelihood differences. The log-likelihood L_{ij} for the data from a given object O_j is the largest when $i = j$, that is, when the parameters have also been estimated using the object O_j data. The other log-likelihoods may be standardized by subtracting this maximum value, L_{jj}. When multiplied by -2, this difference is called the deviance drop.

The deviance drop is:

$$\Lambda_{ij} = -2\{\log[\prod_{l=1}^{m_j} f(y_{jl}; \hat{\mu}_i, \hat{\sigma}_i^2)] - \log[\prod_{l=1}^{m_j} f(y_{jl}; \hat{\mu}_j, \hat{\sigma}_j^2)]\} = -2(L_{ij} - L_{jj}) \qquad i, j = 1 \ldots n$$

Instead of constructing the dissimilarity matrix directly from the deviance drops, the deviance drops may be averaged. This alternative dissimilarity measure, the symmetrized deviance drop, is:

$$D_{ij}^{\Lambda} = (\Lambda_{ij} + \Lambda_{ji}) / 2 \qquad i, j = 1 \ldots n; j < i$$

The deviance drops still exhibit dependence on the number of data points, but it is not clear that this would have a significant impact on the mechanics of the clustering. We will return to this topic in section 4.3. The matrix $\mathbf{D}^{\Lambda}$, with entries D_{ij}^{Λ}, was selected as the measure of dissimilarity to be used in the analyses which follow. $\mathbf{D}^{\Lambda}$ will subsequently be referred to simply as $\mathbf{D}$.

3.5.2 Example - Dissimilarities

The calculation of the quantities in section 3.5.1 is now illustrated for the two objects presented in Table 3.7, O_{36} and O_{12}. Since there are hundreds of claims for each object, the log-likelihoods below are simply expressed as sums of the log-likelihood contributions of each individual observation. Four values are required from the log-likelihood matrix $\mathbf{L}$:

$$L_{36,36} = \log \prod_{l=1}^{m_{36}} f\,(y_{36,l};\hat{\mu}_{36},\hat{\sigma}_{36}^{\,2}) = -\sum_{l=1}^{1475} \{[\log(y_{36,l}) - 7.4063]^2 / [2\,(1.8183)]\}$$
$$-1475 \log[\sqrt{2\pi\,(1.8183)}] - \sum_{l=1}^{1475} \log(y_{36,l}) = -18{,}317.99$$

$$L_{36,12} = \log \prod_{l=1}^{m_{12}} f\,(y_{12,l};\hat{\mu}_{36},\hat{\sigma}_{36}^{\,2}) = -\sum_{l=1}^{3316} \{[\log(y_{12,l}) - 7.4063]^2 / [2\,(1.8183)]\}$$
$$-3316 \log[\sqrt{2\pi\,(1.8183)}] - \sum_{l=1}^{3316} \log(y_{12,l}) = -30{,}264.22$$

$$L_{12,36} = \log \prod_{l=1}^{m_{36}} f\,(y_{36,l};\hat{\mu}_{12},\hat{\sigma}_{12}^{\,2}) = -\sum_{l=1}^{1475} \{[\log(y_{36,l}) - 7.4064]^2 / [2\,(1.8184)]\}$$
$$-1475 \log[\sqrt{2\pi\,(1.8184)}] - \sum_{l=1}^{1475} \log(y_{36,l}) = -18{,}322.63$$

$$L_{12,12} = \log \prod_{l=1}^{m_{12}} f\,(y_{12,l};\hat{\mu}_{12},\hat{\sigma}_{12}^{\,2}) = -\sum_{l=1}^{3316} \{[\log(y_{12,l}) - 7.4064]^2 / [2\,(1.8184)]\}$$
$$-3316 \log[\sqrt{2\pi\,(1.8184)}] - \sum_{l=1}^{3316} \log(y_{12,l}) = -30{,}256.50$$

The deviance drops are then computed directly from the log-likelihood values above:

$$\Lambda_{36,12} = -2\,(L_{36,12} - L_{12,12}) = -2\,[(-30,264.22) - (-30{,}256.50)] = -2\,(-7.72) = 15.44$$

$$\Lambda_{12,36} = -2\,(L_{12,36} - L_{36,36}) = -2\,[(-18{,}322.63) - (-18{,}317.99)] = -2\,(-4.65) = 9.30$$

Finally, the dissimilarity between objects O_{36} and O_{12} is given by the average of these two deviance drops:

$$D_{36,12} = (\Lambda_{36,12} + \Lambda_{12,36}) / 2 = (15.44 + 9.30) / 2 = 24.74 / 2 = 12.37$$

3.5.3 Updating the Dissimilarity Matrix

The dissimilarity matrix must be updated at each step of the clustering algorithm and this will now be demonstrated. Before the first step, **D** has dimension n, which is the number of rows and columns. In the first step, two objects will be merged to form a cluster; then, considering all the other objects as clusters, there will be n-1 clusters in a set G^{n-1} having individual clusters G_i^{n-1}, i = 1…n-1. The dissimilarity matrix involving these clusters will now have dimension n-1. In general, after the k^{th} step, there will be n-k clusters in a set G^{n-k}, and the dissimilarities will be contained in the matrix $\mathbf{D}^{n-k}$ with entries D_{ij}^{n-k}. The updated contiguity matrix will be denoted by $\mathbf{C}^{n-k}$, and have elements C_{ij}^{n-k}.

At the k^{th} step of the algorithm, k = 1…n-1, the next cluster is formed by searching for h_k, the lowest dissimilarity between any two contiguous clusters $G_i^{n-(k-1)}$ and $G_j^{n-(k-1)}$:

$$h_k = \min\{D_{ij}^{n-(k-1)} : C_{ij}^{n-(k-1)} = 1\} \qquad i, j = 1 \ldots n-(k-1);\ j < i$$

Suppose the two clusters satisfying this criterion are $G_{i*}^{n-(k-1)}$ and $G_{j*}^{n-(k-1)}$. The dissimilarity between this newest cluster, $G_{i*}^{n-(k-1)} \cup G_{j*}^{n-(k-1)}$, and all other clusters

$G_g^{n-(k-1)}$, $g \notin \{i_*, j_*\}$, must now be determined. The revised matrix $\mathbf{D}^{n-k}$ is constructed from the preceding dissimilarity matrix $\mathbf{D}^{n-(k-1)}$ by the following steps:

1. Move the two rows and columns of $\mathbf{D}^{n-(k-1)}$ corresponding to clusters $G_{i*}^{n-(k-1)}$ and $G_{j*}^{n-(k-1)}$ to the bottom two rows and the two rightmost columns of $\mathbf{D}^{n-(k-1)}$, thus creating a re-ordered version of the matrix, $\tilde{\mathbf{D}}^{n-(k-1)}$. These two rows and columns will later be replaced with a single new row and column describing the dissimilarities between this newly formed cluster and all the others, which were unaltered in step k.

2. Calculate a vector $\mathbf{w}$ of length n-k, with entries given by (Gordon, 1996b):

$$w_j = \alpha_1 \tilde{D}^{n-(k-1)}_{n-(k-1),\,j} + \alpha_2 \tilde{D}^{n-(k-1)}_{n-k,\,j} + \beta\, \tilde{D}^{n-(k-1)}_{n-(k-1),\,n-k} + \gamma \left| \tilde{D}^{n-(k-1)}_{n-(k-1),\,j} - \tilde{D}^{n-(k-1)}_{n-k,\,j} \right|$$

where the parameters α_1, α_2, β and γ may be varied in accordance with the desired agglomerative algorithm. This general form allows many algorithms to be tested using the same computer program.

Under the complete linkage method, the elements of $\mathbf{w}$ are calculated by setting $(\alpha_1, \alpha_2, \beta, \gamma) = (\frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2})$ in the general agglomerative formula given above. It can be seen that this simplifies to taking maxima of pairs of dissimilarity values:

$$w_j = \max\,[\tilde{D}^{n-(k-1)}_{n-(k-1),\,j}, \tilde{D}^{n-(k-1)}_{n-k,\,j}] \qquad j = 1 \ldots n-(k-2)$$

3. Finally, delete the last two rows and columns in $\tilde{\mathbf{D}}^{n-(k-1)}$ and replace them by the vector $\mathbf{w}$ and its transpose, respectively, making a new last row and new right-hand column. The matrix so formed, of dimension n-k, is the matrix $\mathbf{D}^{n-k}$. The diagonal elements of a dissimilarity matrix are not meaningful, and may be replaced by zeros.

When using complete linkage, at each step of the algorithm, we keep track of the maximum dissimilarities between any two objects in a cluster, and so it is not necessary to re-visit the original dissimilarity matrix $\mathbf{D}$. For the construction of $\mathbf{D}^{n-k}$, all the dissimilarity values required are present in the matrix $\mathbf{D}^{n-(k-1)}$. This type of argument will also apply to single linkage, but not, for example, to average linkage, where the original matrix $\mathbf{D}$ must be consulted at each step.

The contiguity matrix, using complete linkage, can be revised in an analogous way. This is true only because the complete link method is being used. The only difference with the contiguity matrix is that the maxima are taken over values of a binary matrix.

4. Using the updated matrices, clusters are successively agglomerated until only one remains.

3.5.4 Reversals

Overall, complete linkage is the most robust linkage method. Robustness is often assessed in reference to reversals. Suppose two clusters have merged in the previous step. A reversal occurs when a third cluster is more similar to the new cluster than the two clusters which joined are to each other. Conditions necessary for the absence of reversals are given by Gordon (1996a) in terms of the clustering parameters α_1, α_2, β and γ:

(i) $\gamma \geq -\min(\alpha_1, \alpha_2)$;

(ii) $\alpha_1 + \alpha_2 \geq 0$; and

(iii) $\alpha_1 + \alpha_2 + \beta \geq 1$.

Reversals are more likely under contiguity constraints as similar objects may not become contiguous until later in the clustering process, that is, after other intervening objects have joined. In fact, complete linkage is the only major method immune to reversals when contiguity is defined in terms of the individual objects (Murtagh, 1985). This is because condition (iii) above becomes stricter when constraints are present (Gordon, 1996b). The revised condition is:

(iii) $\min[\alpha_1 + \alpha_2, \gamma + \min(\alpha_1, \alpha_2)] + \beta \geq 1$.

3.6 Collision Severity Example - Clustering

The methodology discussed in the previous section will now be applied to collision severity in one particular territory, whose contiguity structure has been determined using the criteria in section 3.2, and whose data have been adjusted as per sections 3.3 and 3.4. Figure 3.1 is the map of this sample territory with 11 objects labelled by letters of the alphabet: A, C, E, J, K, L, M, N, P, S and Z.

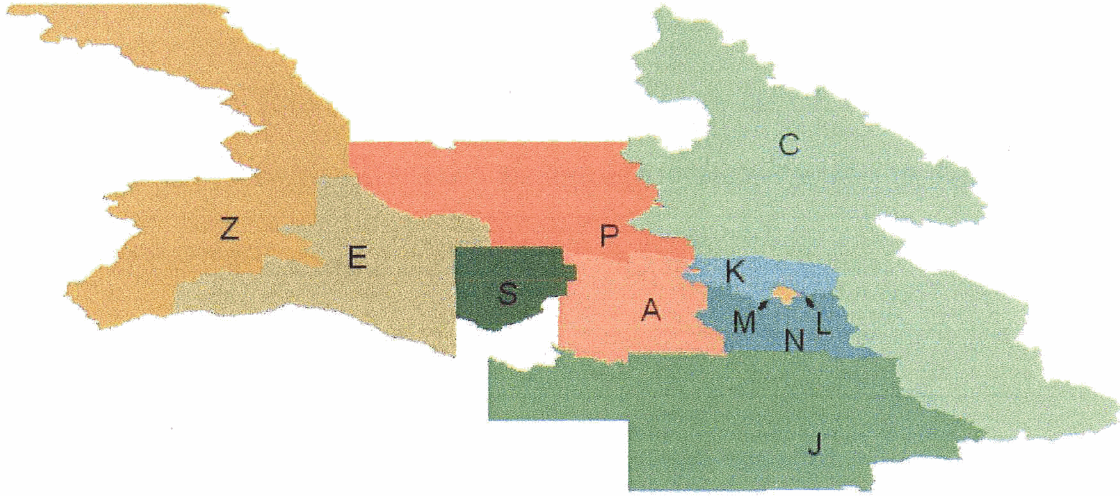

Figure 3.1 Map of original objects in territory

The contiguity matrix for this set of objects is specified in Table 3.8:

	Z	E	S	P	A	C	K	L	M	N	J
Z											
E	1										
S	0	1									
P	0	0	0								
A	0	0	1	1							
C	0	0	0	0	0						
K	0	0	0	0	0	1					
L	0	0	0	0	0	0	1				
M	0	0	0	0	0	0	1	1			
N	0	0	0	0	1	0	1	1	1		
J	0	0	0	0	0	0	0	0	0	1	

Table 3.8 Contiguity matrix $\mathbf{C}$, for selected territory

Although it appears from Figure 3.1 that object P might have more than one neighbour, objects Z, E, S and C are not defined as contiguous for the reasons given in section 3.2.

The matrices in this section are all symmetric and so only the sub-diagonal elements are shown. The object dissimilarity matrix is displayed in Table 3.9, with its values rounded to the nearest integer:

	Z	E	S	P	A	C	K	L	M	N	J
Z											
E	22										
S	63	146									
P	17	55	24								
A	8	5	114	33							
C	2	17	64	13	5						
K	268	94	786	373	158	231					
L	70	16	289	100	28	50	65				
M	239	87	769	297	128	186	56	22			
N	283	67	1137	446	133	231	36	23	32		
J	20	67	183	16	28	7	422	106	287	357	

Table 3.9 Dissimilarity matrix $\mathbf{D}$, for selected territory

To illustrate how clustering proceeds, the 4^{th}, 5^{th} and 6^{th} steps of the complete linkage algorithm are illustrated. After the first three iterations, the contiguity and dissimilarity matrices are as shown in Tables 3.10 and 3.11:

	S	P	A	C	K	J	EZ	LMN
S								
P	0							
A	1	1						
C	0	0	0					
K	0	0	0	1				
J	0	0	0	0	0			
EZ	1	0	0	0	0	0		
LMN	0	0	1	0	1	1	0	

Table 3.10 Contiguity matrix $\mathbf{C}^8$, after 3[rd] step

	S	P	A	C	K	J	EZ	LMN
S								
P	24							
A	114	33						
C	64	13	5					
K	786	373	158	231				
J	183	16	28	7	422			
EZ	146	55	8	17	268	67		
LMN	1137	446	133	231	65	357	283	

Table 3.11 Dissimilarity matrix $\mathbf{D}^8$, after 3[rd] step

The eight smallest values in the dissimilarity matrix above correspond to values of zero in the contiguity matrix, which means that the corresponding mergers are not permitted. It is objects P and A which have the lowest dissimilarity among contiguous pairs. We create a new row at the bottom of the matrices in Tables 3.12 and 3.13 for this new cluster called AP. It is contiguous to object S and cluster LMN since object A was contiguous to both of these, and object P was only contiguous to object A. Notice in the matrices below that the rows and columns corresponding to objects A and P have been deleted.

	S	C	K	J	EZ	LMN	AP
S							
C	0						
K	0	1					
J	0	0	0				
EZ	1	0	0	0			
LMN	0	0	1	1	0		
AP	1	0	0	0	0	1	

Table 3.12 Contiguity matrix $\mathbf{C}^7$, after 4th step

	S	C	K	J	EZ	LMN	AP
S							
C	64						
K	786	231					
J	183	7	422				
EZ	146	17	268	67			
LMN	1137	231	65	357	283		
AP	114	13	373	28	55	446	

Table 3.13 Dissimilarity matrix $\mathbf{D}^7$, after 4th step

The next merger is between a single object, K, and the cluster of three objects, LMN. This is true because neither objects S and C nor clusters EZ and AP are contiguous, and all other lower dissimilarities were previously ruled out. We now give details of the construction of the matrix $\mathbf{D}^6$. Following step 1 of section 3.5.3, we obtain the matrix $\tilde{\mathbf{D}}^7$ displayed in Table 3.14:

	S	C	J	EZ	AP	K	LMN
S							
C	64						
J	183	7					
EZ	146	17	67				
AP	114	13	28	55			
K	786	231	422	268	373		
LMN	1137	231	357	283	446	65	

Table 3.14 Dissimilarity matrix $\tilde{\mathbf{D}}^7$, during 5th step

and the vector $\mathbf{w}$ from step 2 is constructed by taking maxima of the following pairs of values from $\tilde{\mathbf{D}}^7$: (786, 1137); (231, 231); (422, 357); (268, 283); and (373, 446). This vector forms the new row KLMN, which appears at the bottom of Table 3.16. The contiguity matrix $\mathbf{C}^6$ is obtained similarly and shown in Table 3.15.

	S	C	J	EZ	AP	KLMN
S						
C	0					
J	0	0				
EZ	1	0	0			
AP	1	0	0	0		
KLMN	0	1	1	0	1	

Table 3.15 Contiguity matrix $\mathbf{C}^6$, after 5th step

	S	C	J	EZ	AP	KLMN
S						
C	64					
J	183	7				
EZ	146	17	67			
AP	114	13	28	55		
KLMN	1137	231	422	283	446	

Table 3.16 Dissimilarity matrix $\mathbf{D}^6$, after 5th step

In the 6th step, object S joins cluster AP. The matrices for the remaining five clusters are given in Tables 3.17 and 3.18. The last row in Table 3.18 is the transpose of the first column in Table 3.16, after deleting the rows corresponding to object S and cluster AP. This is because cluster AP is less dissimilar to each other cluster than object S is.

	C	J	EZ	KLMN	APS
C					
J	0				
EZ	0	0			
KLMN	1	1	0		
APS	0	0	1	1	

Table 3.17 Contiguity matrix $\mathbf{C}^5$, after 6th step

	C	J	EZ	KLMN	APS
C					
J	7				
EZ	17	67			
KLMN	231	422	283		
APS	64	183	146	1137	

Table 3.18 Dissimilarity matrix $\mathbf{D}^5$, after 6th step

Clustering continues until all objects occupy a single group. The next step would be to combine the two clusters EZ and APS, followed by object C and cluster KLMN, then object J and cluster CKLMN. At the last step, the two remaining clusters are joined.

3.7 Methodology for Frequency Data

We now turn to the discussion of the frequency count data. Unlike the severities, these data are only available in the grouped format of Table 3.19. The number of claims observed, W, during the terms of individual policies belonging to object O_k is tabulated, for all policies having one or more claims. The exposure for each of these policies is also determined. Then the total exposure, E_{kc}, is calculated among all policies having exactly c claims, c > 0, during their terms.

Object (O_k)	**# of claims, c, during term of policy (W)**	**Sum of exposure for policies with exactly c claims during term** (E_{kc})	**Product of # of claims & sum of exposure** ($c\,E_{kc}$)
O_1	0	$E_{10} = E_1 - \sum_{c=1}^{9} E_{1c}$	0
	1	E_{11}	E_{11}
	...	...	...
	9	E_{19}	$9\,E_{19}$
...	...	...	...
O_n	0	$E_{n0} = E_n - \sum_{c=1}^{9} E_{nc}$	0
	1	E_{n1}	E_{n1}
	...	...	...
	9	E_{n9}	$9\,E_{n9}$

Table 3.19 Format of frequency data

For each object O_k, the exposure counts for policies with claims, E_{kc} $(c = 1, 2, \ldots)$ are multiplied by the factor u_k, described in section 3.4. The notation will be unchanged for these adjusted exposures. The total object exposure, E_k, is obtained from a separate database and used to calculate the total adjusted claims-free exposures, E_{k0}, as in Table 3.19.

3.7.1 Collision Frequency

One possible model for these frequencies is the Poisson distribution. The Poisson probability mass function for the number of claims, W, is given by:

$$P(W = c) = p(c; \lambda) = \lambda^c e^{-\lambda} / c! \qquad c \in \{0, 1, 2, \ldots\}$$

The estimate of the rate parameter within each object does not correspond to the usual notion of claims frequency in insurance. The numerator in the formula contains exposures rather than policy counts. This lessens the impact of short-term policies by weighting them with an exposure of less than one. For object O_k, the maximum likelihood estimate of λ_k is easily derived to be:

$$\hat{\lambda}_k = (\textstyle\sum_{c=0,1,\ldots} c\, E_{kc}) / E_k \qquad k \in \{1, \ldots, n\}$$

The log-likelihood of the object O_j data under the model estimated from the object O_i data is:

$$L_{ij} = \log \prod_{c=0,1,\ldots} [p\,(c; \hat{\lambda}_i)]^{E_{jc}} = \log(\hat{\lambda}_i)\,(\sum_{c=0,1,\ldots} c\, E_{jc}) - \hat{\lambda}_i E_j - \sum_{c=0,1,\ldots} E_{jc} \log(c!)$$

As for severities, the dissimilarity matrix consists of the symmetrized deviance drops:

$$D_{ij}^{\Lambda} = (\Lambda_{ij} + \Lambda_{ji}) / 2 \qquad\qquad i, j = 1 \ldots n;\ j < i$$

Clustering then proceeds in exactly the same fashion as for severity data.

3.7.2 Comprehensive Frequency

The Poisson distribution is often used in the context of rare events and thus may not be suitable for modelling the frequency of perils such as vandalism and theft from autos. This is because there are enough policies with a large number of these types of claims to inflate the variance in claim counts far above the mean. The negative binomial distribution is employed in this case.

The negative binomial probability mass function for the number of claims, W, is given by:

$$P\,(W = c) = p\,(c; r, q) = \frac{\Gamma\,(r + c)\, q^c}{c!\,\Gamma\,(r)\,(1 + q)^{r+c}} \qquad\qquad c \in \{0, 1, 2, \ldots\}$$

and so the log-likelihood of the object O_j data using the parameters $\hat{r}_i$ and $\hat{q}_i$ estimated from the object O_i data is then:

$$L_{ij} = \log \prod_{c=0,1,\ldots} [p\,(c;\hat{r}_i,\hat{q}_i)]^{E_{jc}}$$
$$= \sum_{c=0,1,\ldots} E_{jc} [\log\{\Gamma(\hat{r}_i + c) / [c!\,\Gamma(\hat{r}_i)]\} + c \log(\hat{q}_i) + (\hat{r}_i + c) \log(1 + \hat{q}_i)]$$

The symmetrized deviance drop is again used as the dissimilarity measure, and clustering is performed as before.

3.8 Synthesis of Results

Suppose we have the results of the separate cluster analyses on the four coverage and attribute combinations. The desired output is a single set of clusters within the territory. It is not straightforward to determine the best number of clusters. The dissimilarity values are not directly comparable across analyses. One can note jumps in the pattern of values in each case but there is little chance that these will correspond exactly.

Gordon (1999) formalizes the process of combining the results of multiple analyses. He defines three main consensus rules. Under the strict consensus rule, only clusters appearing in all analyses are included as part of the combined set. A somewhat relaxed version of this is the majority consensus rule, which only requires the final clusters to be present in more than half of the individual classifications. Finally, the median consensus rule seeks to minimize the number of classes retained which occur less than half of the time.

The strategy employed here can be described as a combination of consensus and outlier re-allocation, applied in turn. The analyses can be presented in terms of four sets of maps. To have a meaningful comparison between the different results, each clustering sequence must be frozen at the same step of the algorithm and a snapshot taken. Then the maps can be overlaid and a combined set of results produced.

In the Lower Mainland, nearly 3/4 of the objects had at least one neighbouring object which was in the same cluster as the object, for all four combinations of coverage and attribute (see Table 3.20). This was assessed once each of the individual analyses had reached 11 clusters, a number determined largely by trial and error. For the Southern Interior, just under 2/3 of the objects were in perfect agreement, while the percentage was higher in the medium-sized territories.

Territory	**Perfect Matches**
Kootenays	92%
Mid Island/Sunshine Coast	91%
Prince George Area	82%
Lower Mainland & Fraser Valley	72%
Southern Interior	64%

Table 3.20 Proportions of objects satisfying strict consensus rule

This process leaves a minority of the objects unassigned to a cluster. Judgement is then applied to fill in the gaps. There are several considerations which can be helpful:

(i) a singleton object could be landlocked by a cluster;

(ii) an object could match a particular cluster on three of the four maps and it seemed reasonable to accept the single difference;

(iii) a group of outstanding objects could be formed into an additional cluster because they were most similar to each other;

(iv) one could determine that an object was about to join a cluster by examining what the next few mergers would have been.

By using these considerations, a reasonable set of consensus clusters was produced in each existing territory. The cut-off number of steps was chosen for each territory having in mind an acceptable range for the number of new clusters, based on the population of the original territory.

3.8.1 Example - Synthesis

We now show how a set of results obtained as in section 3.6 can be combined. In the example, the first two mergers for collision severity, as well as the second two, occurred at very comparable dissimilarity values. After that point, the gaps gradually increased, until there was a very large jump in dissimilarity when performing the final merger. Table 3.21 displays the dissimilarity values of the mergers for all four analyses run on this territory.

Clusters Remaining	Collision Severity	Comp. Severity	Collision Frequency	Comp. Frequency
10	22	7	0	7
9	22	23	1	9
8	32	37	3	27
7	33	93	5	29
6	65	97	20	39
5	114	168	26	136
4	146	179	38	416
3	231	306	59	753
2	422	423	75	763
1	1137	1556	88	1525

Table 3.21 Dissimilarity values for mergers

The results of the four cluster analyses must then be synthesized. The information in Table 3.21 is used as an aid in deciding where clustering should be stopped. We observe for comprehensive frequency, it would not be advisable to stop at three clusters, as another merger could be performed with a minimal increase in dissimilarity. Since there are only 11 objects in the territory, stopping with three or more clusters would likely result in too many groupings. Therefore, a snapshot of the four analyses was taken with two clusters remaining. Four maps were produced, and when overlaid, resulted in the groupings in Figure 3.2.

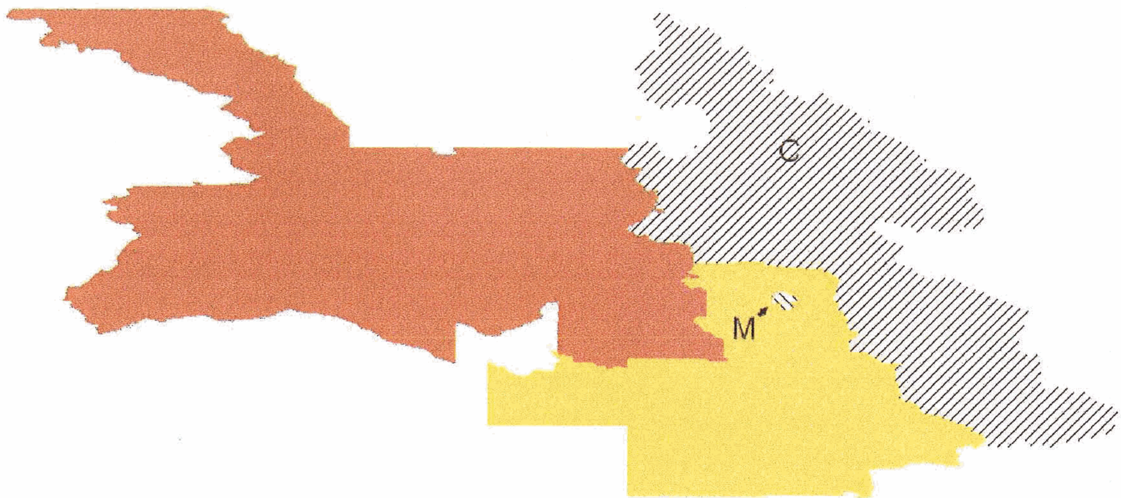

Figure 3.2 Strict consensus clusters

Nine of the 11 objects from Figure 3.2 were in agreement on all four maps. The other two hatched objects, C and M, did not satisfy the strict consensus rule. Object C matched with the cluster JKLN on three of the four maps, while in case of comprehensive severity, it stood alone as a cluster and all other objects occupied the other cluster. The small circular object, M, also matched the cluster JKLN on three of the four maps, but not the same three maps as for object C. It was its own cluster for collision frequency. If the majority consensus rule was applied, objects C and M joined cluster JKLN to produce the two clusters displayed in Figure 3.3.

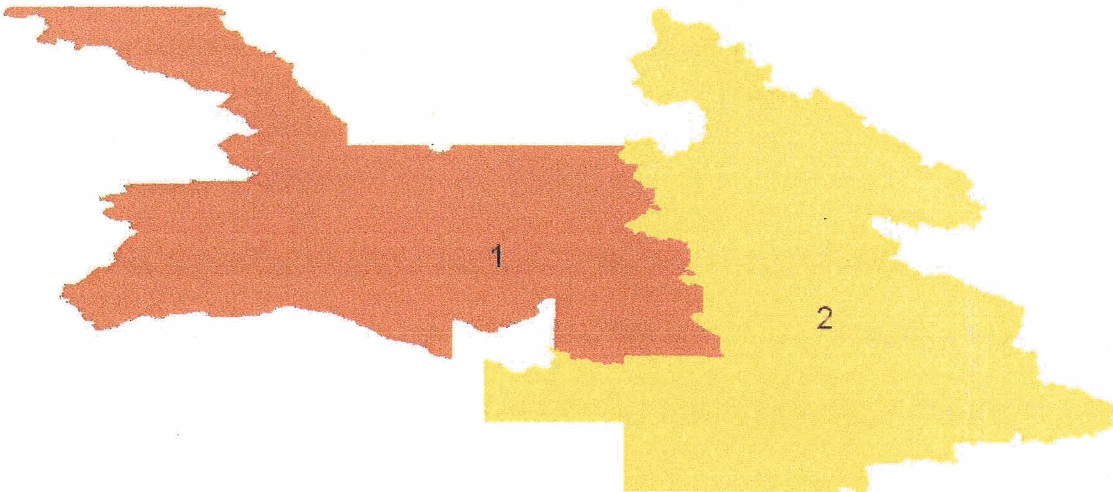

Figure 3.3 Majority consensus clusters

3.9 Outlying Objects

Upon inspection of the object loss ratios within the clusters in each territory, a number of objects was identified which did not seem to fit well in their assigned clusters. It was then decided to allow these areas to move to clusters that were not necessarily contiguous but contained objects with more similar loss ratios. In cases where a better suited match was not found, no change was made.

The decision to redistribute some objects was generally based on outlying loss ratios. It was then of interest to compare how the adjusted loss costs fit in after these re-groupings. The adjustment refers to isolating the territory effect by controlling for other rating factors, as discussed in sections 3.3 and 3.4. It turns out that most of the differences between loss cost and loss ratio relativities were less than 3%. This would indicate that the data normalization process had been performed adequately.

3.9.1 Example - Outliers

At this point, the loss ratios were examined for the objects in each cluster to ensure that any outlying objects were re-assigned. Object C was identified as an outlier, as its loss ratios were somewhat higher than those of the other objects in cluster 2. The decision was made to re-allocate object C to cluster 1, resulting in the revised territory map in Figure 3.4.

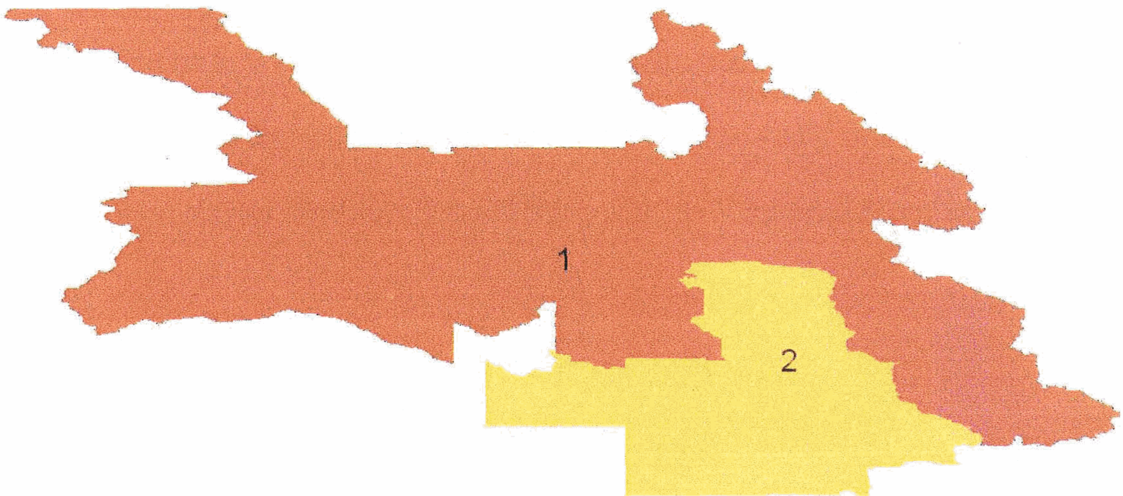

Figure 3.4 Map of final clusters for territory

3.10 Other Territories

The clustering algorithms were not run on any of the rural territories in northern British Columbia or smaller territories along the coast. Due to their small populations, it would not have been advisable to split these territories.

The only urban territory which did not produce satisfactory results was the Victoria area. This territory incorporates a strange mixture of regions including Southern Vancouver Island as well as other islands west of the mainland. It includes relatively accessible areas, and also less populated and more remote regions like the Queen Charlotte Islands.

A major difficulty lay in defining contiguity among islands. Many of the islands are connected by ferry to ports located in different territories. This means the island is not contiguous to any other part of its own territory, except possibly other islands. An additional problem is the lack of data, as most of these islands have very small populations. It was decided to form four island groups that would have a credible amount of data. These objects were not clustered along with the Vancouver Island portion of the territory.

4 Diagnostics

4.1 Alternative Models for Severity

The lognormal distribution used in chapter 3 is just one possible model for claims severities; one might also consider Gamma, Weibull, Pareto, exponential and other distributions (Klugman et al., 1998). Some of the other common loss distributions were not considered for various reasons: the Pareto distribution is not available in standard statistical computing packages; normal and logistic distributions are symmetric, while the data presented here are clearly not; Weibull, inverse distributions and various transformations were considered unnecessarily complex. We consider only the lognormal and Gamma families.

4.1.1 Collision Severity Models

Plots were constructed to compare the cumulative distribution functions of the fitted lognormal models to the empirical CDFs of collision severities. The plots showed that the lognormal distribution does not generally fit well in the tails. Figure 4.1 is the plot of the empirical (solid) and fitted (dotted) distributions for one such representative object. The scale along the x-axis has been removed for confidentiality reasons.

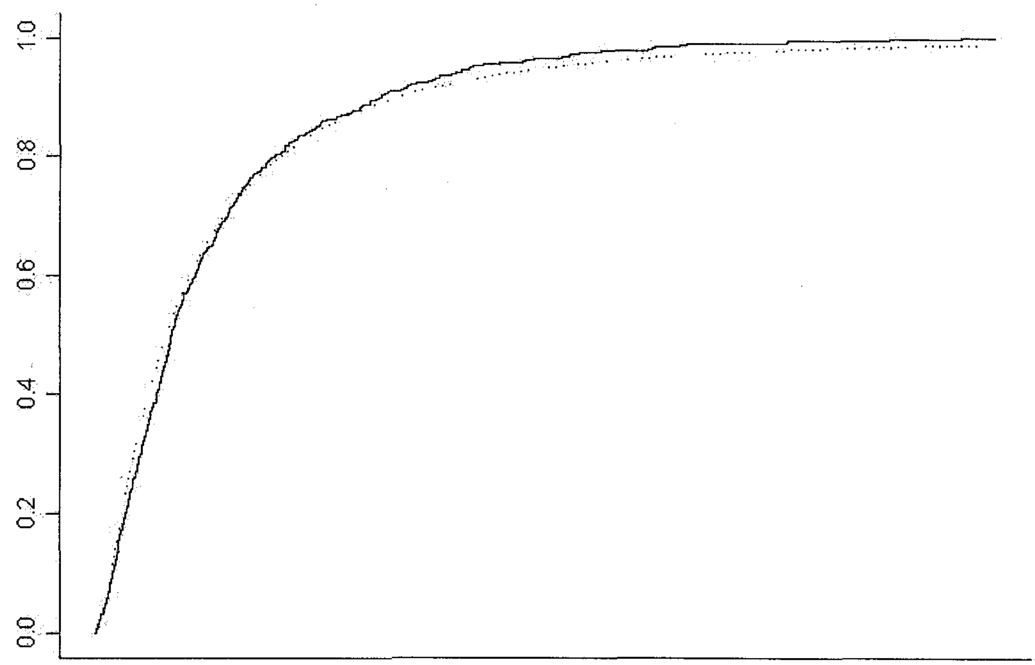

Figure 4.1 Empirical and hypothesized lognormal CDFs

The two parameter gamma distribution is a much better fit in the upper tail, however it is not a good fit in the lower tail. This is likely because method-of-moments estimators were used, and these are greatly affected by larger values. The exponential distribution was examined and fits quite well, especially in the tails. This suggests that the density may be strictly decreasing, instead of initially increasing. However, upon inspection of the empirical distribution functions, this monotonicity was not observed for all of the objects. The collision severity cluster analysis based on an exponential model was run for one territory, but it resulted in one of the clusters containing a disproportionate number of objects. Finally, the best option seemed to be to tolerate the imperfections of the lognormal model, assuming that it might not affect the clustering outcomes greatly.

4.1.2 Comprehensive Severity Models

The lognormal model seemed to perform much better for comprehensive severities than collision severities, especially in the tails. The exponential model is not a good fit here. Gamma distributions fit well only in the right tails, again likely due to the chosen method of estimation. Furthermore, the lognormal distribution seems to fit adequately based on the plots of the empirical and estimated distribution functions.

4.2 Alternative Models for Frequency

Three counting distributions, the binomial, Poisson, and negative binomial, are commonly used in insurance (Klugman et. al, 1998). The binomial distribution is useful in life insurance, or other instances where at most one claim occurs for each of a set of policies, and where the variance in claim counts does not exceed the mean. For collision and comprehensive automobile insurance coverages, multiple claims are possible and variances are not generally less than means. Therefore, we consider only the Poisson and negative binomial distributions.

4.2.1 Collision Frequency Models

The Poisson is a single parameter counting distribution with the restriction that the mean equals the variance. The results of the χ^2 test for the Poisson distribution in the territory with the highest collision frequency showed one quarter of the statistics significant at the 1% level. This was because the probabilities of two or more claims were being underestimated by the Poisson model. The estimates of the overdispersion were mostly in the 1 - 3% range. However, when attempting to fit negative binomial models for collision frequency, there were several objects with a maximum observed value of two claims during a policy term, and for these objects, the variance was not greater than the mean. It was expected that an even greater problem would arise with such objects in the other territories with lower collision frequencies. Therefore, the Poisson distribution was selected to model collision frequency.

4.2.2 Comprehensive Frequency Models

Comprehensive frequency has a much longer-tailed distribution than collision frequency since occurrences of multiple claims are more likely. However, there are no claims against the majority of policies during a year. Most of the probability mass function for comprehensive frequency is still concentrated at 0 and 1, as it is for collision frequency. Then fitted Poisson models have tails which decrease too quickly to account for the probabilities of five or more comprehensive claims. Policies with this many comprehensive claims were present in most objects, and Poisson models were easily rejected by the χ^2 test.

Since the variances in comprehensive frequency clearly exceed the means, a negative binomial model is proposed. For one of the territories with moderate frequencies, the likelihood ratio tests of Poisson versus negative binomial favoured the latter distribution in every case. In this territory, about 10% of the objects were not a good fit to the negative binomial model, according to the χ^2 test at a 1% significance level. The negative binomial model would be rejected for one-third of the objects if the significance level were set at 10%.

4.3 Dissimilarities

A number of graphs were created to check various properties of the dissimilarities during the clustering of Lower Mainland collision severity. The iteration number for the first merger of each object, and the corresponding dissimilarity value, were plotted against both the total number of claims, and the number of original neighbours. Also, mean severity and its rank among all values were plotted against the time of the first merger. None of these graphs indicated any obvious biases in the complete linkage algorithm.

Three of the charts described above are presented here. Each point in the graphs represents an individual object at the time it first joins a cluster. Figure 4.2 shows the dissimilarity values associated with the first mergers of the original objects in a territory. In the later iterations, it does not appear that the size of an object influences when it first merges. Apart from the outlier with the largest dissimilarity, if one examines the sizes of objects across any range of dissimilarity values above 20, the distribution of object size does not appear to depart significantly from a uniform distribution.

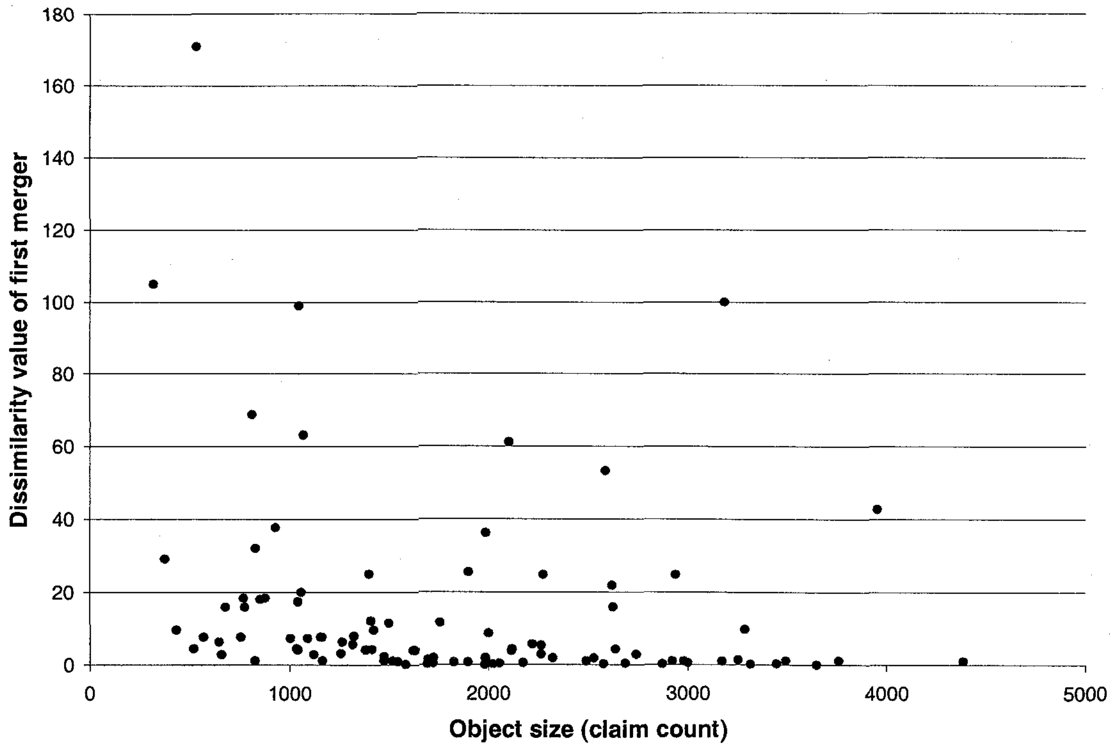

Figure 4.2 Dissimilarity values of first mergers

In Figure 4.3, the mean severity for each object is plotted as clustering proceeds in time along the x-axis. When there are two points at a single step, it indicates that two singleton objects were combined to form a new cluster. This is the case in most of the early iterations. Later on, a single severity value for a step means that a singleton object joined an existing cluster. If there is no point plotted for a particular iteration, two existing clusters were merged at that step. This is common in the later stages of agglomeration, after most objects have joined a cluster, which occurs by about step 45 in Figure 4.3.

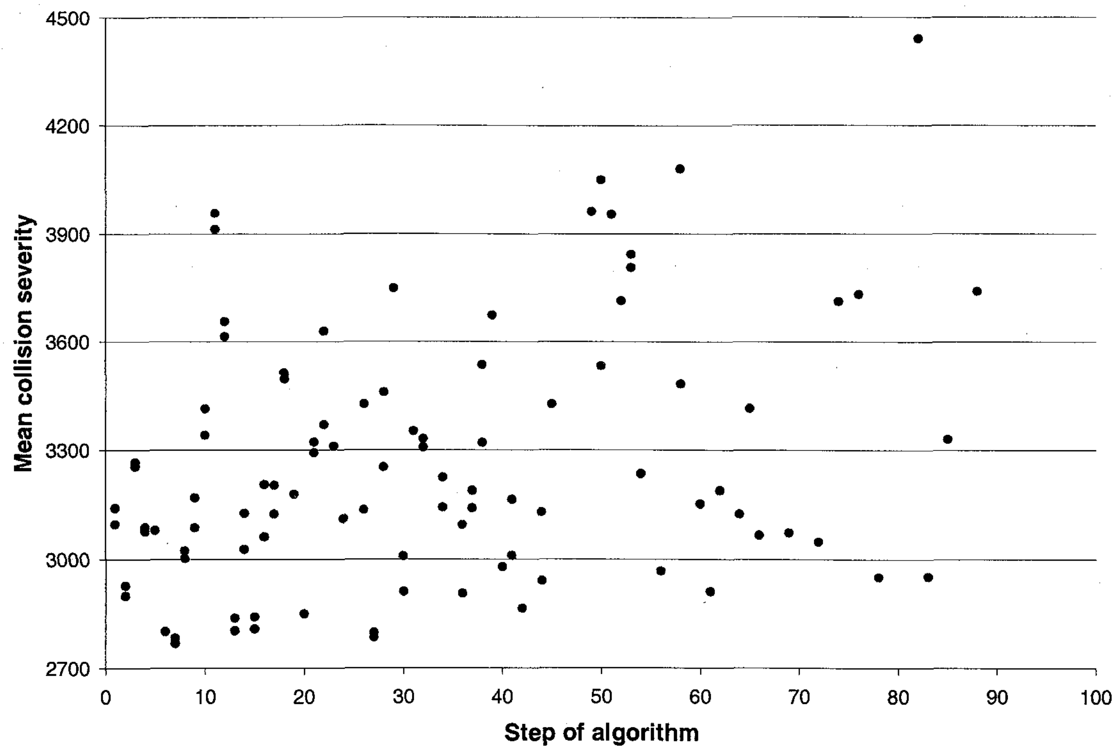

Figure 4.3 Collision severities of merged objects

Figure 4.4 reveals the relationship between the number of neighbours an object has in the original contiguity matrix, and how soon it first joins a cluster. It might seem plausible that an object with more neighbours would be more likely to join a cluster earlier. However, for objects with any given number of original neighbours, there are always some objects which merge early and others which merge late in the algorithm.

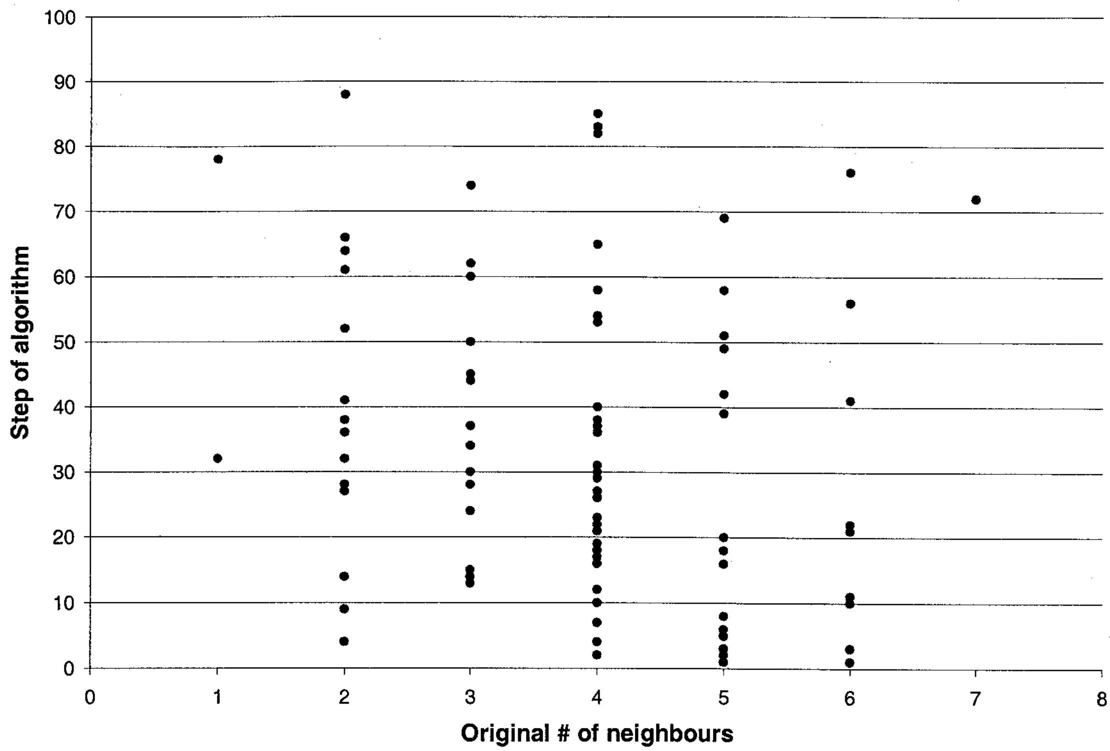

Figure 4.4 Steps of first mergers by number of neighbours

One final graph was created to assess the dependence of the symmetrized deviance drops on object size. The dissimilarity value upon merger of the last few singleton objects to join a cluster is such that it ranks between the 70th and 85th percentile of all values in the original dissimilarity matrix $\mathbf{D}$. Figure 4.5 does show an increasing trend in the magnitude of 80th percentiles of object dissimilarities, $d_{.80}$, versus object claim counts. These percentiles were determined by ordering the n-1 dissimilarities between a given object and all others.

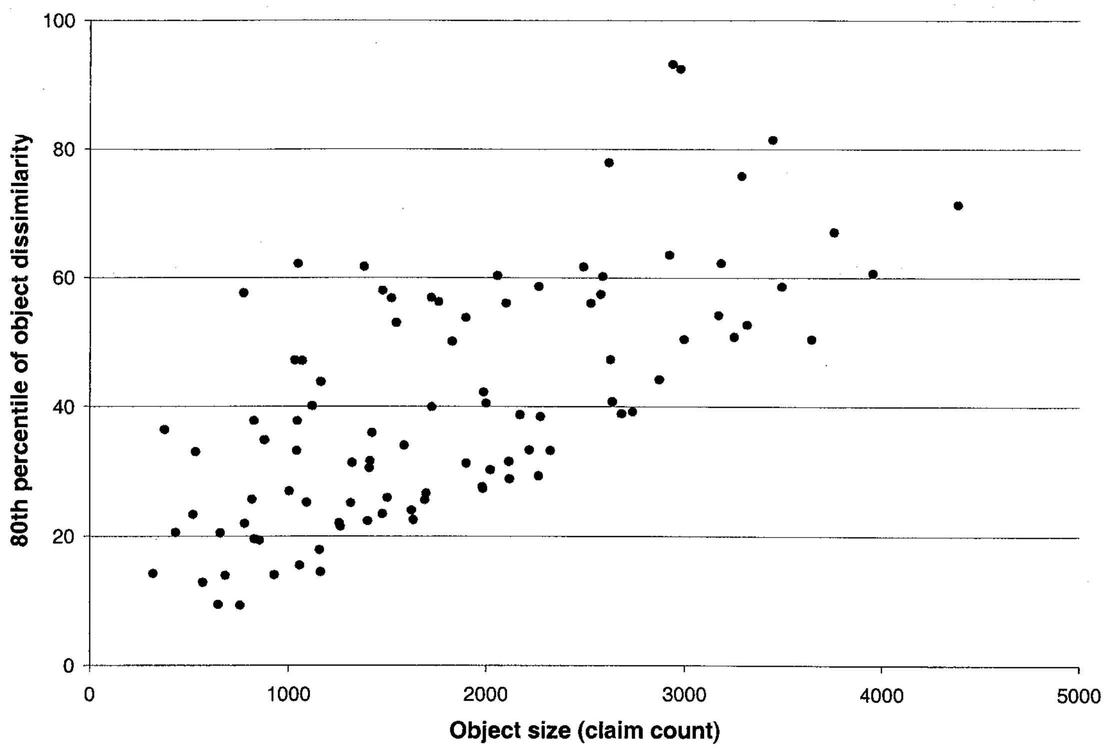

Figure 4.5 80th percentiles of object dissimilarities by number of claims

It should be noted that these higher percentiles of the dissimilarity values only affect the clustering for the few objects which do not merge until late in the algorithm. It may be that most of the clustering process is not greatly affected by the dependence on the number of claims. This is an issue which could be more fully explored in the future.

4.4 Other Clustering Methods and Algorithms

4.4.1 Single Linkage

For two reasons, the single linkage method was not as compatible with the objectives of this study as the complete linkage method. First of all, single linkage is prone to chaining. It would not be desirable here to have one cluster which contained far more objects than the other clusters. Secondly, single linkage produces very differentiated clusters, but they may not be very homogeneous. It is important that objects be similar to others in the same cluster, since the part of the insurance rate which is based on the territory will become the same for all objects in the cluster. It is not a problem if a separate cluster happens to have similar characteristics. A single linkage algorithm was tested on Lower Mainland collision severity data. In the intermediate iterations, a steady 70 - 80% of the clusters were singleton objects, meaning that chaining did take place. This reinforced our choice of the complete linkage method.

4.4.2 Contiguity Constraints

It was necessary to write specialized computer programs to ensure contiguity of clusters. This is because clustering functions and procedures in the standard statistical computing packages do not handle geographical contiguity constraints. The contiguity structure would need to be incorporated directly into the dissimilarity matrix in order to make use of existing software. It seemed unlikely that actual data differences would be preserved under this approach so it was not attempted.

Also, an unconstrained analysis was produced for comparison purposes, however the agreement between it and the corresponding constrained classification was only 11%. In other words, only one in nine objects belonged to the same cluster under both approaches. The removal of contiguity constraints was soon dismissed since this type of analysis would rely heavily on tedious judgements to determine the final clusters.

It is instructive to note that when a set of three territories was clustered simultaneously, the resulting clusters followed the boundaries of the original territories without any overlap. This satisfied the legal constraint on the analysis, which would not likely be met if the entire province were clustered at once.

5 Future Studies

Many aspects of this cluster analysis could benefit from further investigation. Future work may be grouped into three main categories, those issues which would affect the analyses prior to, during and after clustering, respectively:

(i) data used in clustering: object definitions, alternate attributes, sources of data variation;

(ii) dissimilarity and contiguity matrices: dissimilarity measures, distributional models, and approaches to contiguity;

(iii) post-analysis validation: alternatives to consensus clusters, methods of evaluating and improving homogeneity and stability of clusters.

The next three sections describe various ideas belonging to each of these categories.

5.1 Additional Data

Competitors are likely to respond to any change in pricing strategy, so it will be important at some point after implementation of new territories to study the variation within objects (i.e. better and worse neighbourhoods). The postal code is probably too fine a geographical breakdown but the vital statistics code or census tract might work well if they could be mapped from the existing structure. If a finer breakdown than the current proposal is ultimately employed, using surrounding clusters to develop rates would become more important, as less of the individual objects would possess sufficient claims histories to accurately estimate their true claims characteristics.

In some ways, loss cost is a better claims attribute to look at than frequency or severity alone as it incorporates both components of the loss process. However, loss cost is generally a single number defined for a group of policies and thus there is no basis for comparison between groups other than absolute differences. (The same could be said for group frequency or mean severity). To get a meaningful comparison, loss costs are needed on an individual policy basis or other finer breakdown, and these are difficult to obtain. The problem with individual policy loss costs is that they are not uniformly based on annual policy terms. This causes difficulty in calculating the number of policies with no claims, and in using the shorter term data meaningfully. We could also use a grouping such as monthly or annual loss costs to compare objects but it is not clear how dissimilarity would be measured in these cases.

Eventually, leased vehicles could be included in the study as geographical information becomes available; this could result in more accurate groupings in the future. Some further adjustments to the data might also improve the clustering results. Losses could be modelled including deductibles, and data could be adjusted for historical trends where frequencies have changed due to loss reduction initiatives. Another source of data variation which might be adjusted for is short-term policies.

5.2 *Mathematical Issues*

A major question which is challenging to assess is the effectiveness of the dissimilarity measure, e.g. would a sample size adjustment to yield an asymptotic χ^2 distribution for the deviance drops produce better results? The use of a non-parametric dissimilarity measure, such as the difference between the Kolmogorov-Smirnov statistics for two empirical frequency or severity distributions, could also be considered.

With a moderate amount of effort, it would be possible to fit a number of additional loss distributions, besides those of the lognormal and Gamma families, to the severity data. This could reveal if a more complex model would give significantly different or better clustering results.

A slightly different approach to the clustering algorithm would be to require contiguity up to a certain step, and then to begin to combine scattered groupings. This would not likely produce contiguous final clusters, and would require a change in historical direction to be deemed acceptable.

5.3 *Cluster Assessment*

If separate final clusters were allowed for collision and comprehensive coverages, further simplifications to the groupings would be possible. For example in the Lower Mainland, most of Vancouver had similar collision results but there was a large variation in comprehensive experience. This strategy would also help to deal with some problem objects which matched their assigned cluster for one coverage but not the other.

Social census data such as income, education and population could be incorporated to form more homogeneous clusters. It would be difficult to include this data directly in the clustering process, so it would likely be simpler to integrate it when choosing the number of clusters, or allocating outliers to clusters.

A final topic to explore would be splitting the data by time period to assess the stability of territory assignments. It is suspected that the groupings would not change greatly, based on the fact that the loss ratio relativities of each cluster changed very little when the data were updated to include 2002 claims. The differences were less than 3% for the majority of clusters.

Bibliography

Bengio, Y. (2001). Personal communication. University of Montreal.

Carvalho, M. S., Cruz, O. G. and Nobre, F. F. (1996). Spatial Partitioning using Multivariate Cluster Analysis and a Contiguity Algorithm. *Statistics in Medicine* **15**, 1885-1894.

DeSarbo, W. S. and Mahajan, V. (1984). Constrained Classification: The Use of A Priori Information in Cluster Analysis. *Psychometrika* **49**, 187-215.

Flynn-Thapalia, C. (1998). *Multivariate Statistics: Cluster Analysis.* Retrieved February 27, 2001 from http://trochim.human.cornell.edu/tutorial/flynn/cluster.htm.

Gordon, A. D. (1999). *Classification.* Chapman & Hall/CRC, Boca Raton, Florida.

Gordon, A. D. (1996a). Hierarchical Classification. In *Clustering and Classification* (Arabie, P., Hubert, L.J. and De Soete, G., eds.), pages 65-121. World Scientific Publishing Co., Inc., Singapore.

Gordon, A. D. (1996b). A Survey of Constrained Classification. *Computational Statistics and Data Analysis* **21**, 17-29.

Gordon, A. D. (1987). A Review of Hierarchical Classification. *Journal of the Royal Statistical Society, Series A* **150**, 119-137.

Gordon, A. D. (1973). Classification in the Presence of Constraints. *Biometrics* **29**, 821-827.

Hartigan, J. A. (1975). *Clustering Algorithms.* John Wiley & Sons, Inc., New York.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, Inc., New York.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (1998). *Loss Models: From Data to Decisions.* John Wiley & Sons, Inc., New York.

Kolsky, J. D. (1997). *Interpreting Results from Cluster Analysis.* Retrieved February 23, 2001 from http://www.infosens.com/publications/papers/clusteranalysis.html.

Mirkin, B. G. (1996). *Mathematical Classification and Clustering.* Kluwer Academic Publishers, Dordrecht, Netherlands.

Murtagh, F. (1985). A Survey of Algorithms for Contiguity-Constrained Clustering and Related Problems. *The Computer Journal* **28**, 82-88.

Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. Robert E. Krieger Publishing Co., Malabar, Florida.

Smyth, P. (1997). Hierarchical Classification. In *Advances in Neural Information Processing* **9** (Mozer, M. C., Jordan, M. I., and Petsche, T., eds.), pages 648-654. MIT Press, Boston.

Stockburger, D. W. (1996). *Multivariate Statistics: Concepts, Models, and Applications*. Southwest Missouri State University, Springfield, Missouri. Retrieved February, 2001 from http://www.psychstat.smsu.edu/multibook/mlt04m.html.

Walker, M. D. (1998). *Cluster Analysis*. Retrieved February, 2001 from http://www.colorado.edu/epob/epob640mwalker/lect9.html.

Wulder, M.A. (n.d.). *Cluster Analysis*. Retrieved February 23, 2001 from http://www.pfc.cfs.nrcan.gc.ca/landscape/inventory/wulder/mvstats/cluster.html.